Safety Criteria for Selecting a Smart Corridor: Random Forest Approach using HSIS Data from Washington State

Xiaoyu Guo
Email: xiaoyuguo@tamu.edu

Yongxin Peng
and

Chaolun Ma

Zachry Department of Civil and Environmental Engineering
Texas A&M University
College Station, Texas 77843-3136
ABSTRACT

Considering the safety benefit brought by the smart corridor and the better driving experience the smart corridor provides, it is crucial to select the right corridor for the deployment. Moreover, it is important to choose the right site efficiently and effectively, as constructing and maintaining a smart corridor is costly and is a technology challenge. This study demonstrated a use of the HSIS dataset to determine key safety criteria for selecting a smart corridor using a machine learning approach, Random Forest. The authors finalized 13 key safety criteria for selecting a smart corridor out of 111 variables in the HSIS and the HPMS from the State of Washington: (1) Number of crashes, (2) Number of driveway crashes, (3) Number of Intersection crashes, (4) Number of multi-vehicle crashes, (5) Number of crashes in AM Peak, (6) Number of crashes in PM Peak, (7) Number of crashes with property damage only, (8) Left shoulder width, (9) Curve with 28 degree or over, (10) Median width, (11) Average Annual Daily Traffic (AADT), (12) Number of intersection, and (13) Number of signalized intersection. Then, through the evaluation of those key criteria with its existing SPaT enabled corridor, the authors believe that those criteria are critical to consider when selecting a smart corridor, as safety issues such as large crash volumes on a signalized high-speed corridor demonstrates the room for the need of V2I applications. These criteria also agree with the guideline from the FHWA & the NOCoE, when choosing a SPaT enabled corridor is to reduce crashes and improve safety. Lastly, this study predicted potential smart corridors on WA 161, WA 99, WA 202 and discussed their potentials in deploying connected technologies by comparing to the existing smart corridor on WA 522. The key criteria recommended in this study for selecting a smart corridor is generalized and ready to adapt in other states.
INTRODUCTION

A roadway crash is a multifaceted event involving circumstances such as highway geometry, traffic exposure, operating speed, driver characteristics, vehicle factors, as well as the interactions among them. The determination of the relationships between vehicle operating speed, roadway design elements, and traffic volume on crash outcomes has been considered as a goal that would greatly benefit the road safety profession in general. There is a need and an increasing trend to use data-driven procedures, such as machine learning approaches, artificial intelligence, logistic regression methods, databases, such as the Highway Safety Information System (HSIS), contain quality data on a large number of crashes and their associated roadway and traffic records consistently across multiple years and states. These databases provide solid resources to perform innovative learnings.

With the raise of the Intelligent Transportation Systems (ITS) and implementations of technologies in vehicles and infrastructures, various types of detectors, sensors, and cameras have started to be installed in our cars and roadway systems. The ultimate aim of deploying technologies is to reduce crashes, improve safety and achieve Vision Zero. Before massively implementing technologies, a smart corridor, a testbed or a pilot site is an effectively way to deploy technologies and test the impacts brought by them. Most of those on-going smart corridors are constructed to practice and challenge the technology deployments, especially the Vehicle-to-Infrastructure (V2I) communications, for instance, the North Avenue Smart Corridor launched by the City of Atlanta and Georgia Tech. Moreover, there are pilot sites with a larger scale, which implement Vehicle-to-Everything (V2X) on the top of V2I. For instance, the Connected Vehicle Pilot Deployment Program in New York City, Tampa and Wyoming supported by the U.S. Department of Transportation (U.S. DOT). Although there are smart corridors built for both interstates, state highway and urban streets, researchers have revealed that the frequency of crashes was higher, when highways pass through the vicinity of major cities pointing to heavy vehicular movement. Hence, in this study, the selection range of smart corridor is on state highways only.

Recently, the American Association of State Highway and Transportation Officials (AASHTO) initialed the Signal Phase and Timing (SPaT) Challenge and promoted it through National Operations Center of Excellent (NOCoE). A guideline along with the challenge suggested the state DOT and cities to involve at least two high level types of decisions when selecting a SPaT enabled corridor: (1) Need for V2I applications; and (2) Infrastructure compatibility. However, there are more than these two decisions. The selection process is not only complex as suggested in the guideline, but also contains multitudinous options (i.e., potential routes). For instance, in the State of Washington, if an agency only considers state route, then there are 221 options; if an agency applied an additional filter over the length (i.e., in between 22 and 26 miles), then there are still over 80 adequate options. Moreover, the cost of constructing a smart corridor can indeed be expensive. The installation, deployment and maintenance of equipment such as radar, camera and roadside units are to what extent expensive. Thus, it is important to choose the right site efficiently and effectively, as constructing and maintaining a smart corridor is costly, and is a technology challenge.

Although there are existing frameworks, such as the Prioritization Criteria and Methodology Chapter in the Arterial Smart Corridor Projects Final Report, they conducted purely from the state of the practice. To the best of authors’ knowledge, there is not yet any work
to determine important safety criteria for selecting a smart corridor via any machine learning approach. Random Forest (RF) is one of the well-known machine learning techniques for building multiple decision trees and merging them together to obtain a more accurate and stable prediction. RF is also widely applied as it is a good indicator of the importance assigning to the features. In this study, the authors implement a RF model to identify 13 safety criteria out of 111 variables in the HSIS and the Highway Performance Monitoring System (HPMS) data from the State of Washington for selecting a smart corridor. Then, the study evaluates those criteria with its existing SPaT enabled corridor on WA 522. Lastly, this study predicts four potential smart corridors on WA 161, WA 99, and WA 202, and discusses on their potentials in deploying connected technologies. The 13 criteria recommended in this study for selecting a smart corridor is generalized and ready to adapt in other states. As the selection process of a smart corridor is time-consuming and the costs of it is high, the recommended criteria are efficient and effective ways for state and local agencies to identify potential smart corridors in their state route network.

**DATA DESCRIPTION**

Data for the analyses in this study are composed of the HSIS (accident-based) and the HPMS (roadway-based) database in the State of Washington during 2015. The HSIS is a database funded by the U.S. Federal Highway Administration (FHWA). Safety researchers have widely used the database to investigate various topics ranging from problem-identification, modeling to crash-prevention, and prediction [23]. Different from the conventional use of the HSIS data, we aim to identify the safety factors could be used in the selection process of a smart corridor for the implementing ITS related technologies and deploying connected and autonomous vehicles. With the HSIS database as the main source of data, the HPMS is a supportive database that includes data on the extent, condition, performance use and operating characteristics of the nation’s highways. The HPMS data is a roadway-based (or segment-based) data frame, which means each row is one segment in the road network. Thus, the authors integrated the HSIS data with the HPMS data based on route ID and the milepost [24].

As one in the first group of state agencies that undertook the SPaT Challenge, WSDOT is assumed to choose the SPaT corridor on WA 522 by considering various transportation aspects (i.e., safety issues, traffic congestions) and carefully following the guideline [20]. With this assumption, the authors developed a RF algorithm to determine safety criteria for corridor selection process. The RF algorithm was developed by using the data associated with those selected variables (see the Step 2 selection process in the next section, Two-Step Criteria Method) on WA 522. 75 percent in the dataset is randomly sampled as the training dataset, and the rest as the test data. A descriptive statistics summary table of those selected variables are in Table 1.

<table>
<thead>
<tr>
<th>Table 1. Descriptive Statistics Summary on Selected Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspect</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>Accident</td>
</tr>
<tr>
<td>Accidents</td>
</tr>
</tbody>
</table>

1 = Property Damage Only, 2 = Injury, 3 = Fatal
<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accident Type</td>
<td>Categorical</td>
<td>1 = Multi-Vehicle, 2 = Single Vehicle, 3 = Pedestrian/Bike, 4 = Others</td>
</tr>
<tr>
<td>Accident Location</td>
<td>Categorical</td>
<td>1 = Intersection, 2 = Driveway, 3 = Others</td>
</tr>
<tr>
<td>Time of Accident</td>
<td>Categorical</td>
<td>1 = AM Peak, 2 = PM Peak, 3 = Off Peak</td>
</tr>
<tr>
<td>Right Shoulder Width</td>
<td>Numerical</td>
<td>0.0 - 22.0, 1.3 - 2.1</td>
</tr>
<tr>
<td>Left Shoulder Width</td>
<td>Numerical</td>
<td>0.0 - 40.0, 6.0 - 4.2</td>
</tr>
<tr>
<td>Lane Width</td>
<td>Numerical</td>
<td>8.0 - 18.0, 11.8 - 0.6</td>
</tr>
<tr>
<td>Median Width</td>
<td>Numerical</td>
<td>0.0 - 99.0, 20.2 - 30.6</td>
</tr>
<tr>
<td>Road Inventory Grade</td>
<td>Categorical</td>
<td>1 = 0.0 – 0.4 percent, 2 = 0.4 – 2.4 percent, 3 = 2.5 - 4.4 percent, 4 = 4.5 - 6.4 percent, 5 = 6.5 - 8.4 percent, 5 = 8.5 or greater</td>
</tr>
<tr>
<td>Curve</td>
<td>Categorical</td>
<td>1 = Under 3.5 degrees, 2 = 3.5 - 5.4 degrees, 3 = 5.5 - 8.4 degrees, 4 = 8.5 – 13.9 degrees, 5 = 14.3 – 27.9 degrees, 6 = 28 degrees or more</td>
</tr>
<tr>
<td>Traffic AADT</td>
<td>Numerical</td>
<td>159 - 237,647, 40,402 - 49,963</td>
</tr>
<tr>
<td>Traffic AADT Single Truck</td>
<td>Numerical</td>
<td>5 - 7,550, 1,294 - 1,448</td>
</tr>
<tr>
<td>Traffic Number of Signalized Intersection</td>
<td>Numerical</td>
<td>0.0 - 9.0, 0.6 - 1.4</td>
</tr>
<tr>
<td>Traffic Number of Intersection</td>
<td>Numerical</td>
<td>0.0 - 78.0, 3.6 - 5.7</td>
</tr>
<tr>
<td>Traffic Percent Peak Single-Unit Trucks and Buses</td>
<td>Numerical</td>
<td>0.0 - 4.0, 0.3 - 0.2</td>
</tr>
</tbody>
</table>

Note: min. = minimum; max. = maximum; S.D. = standard deviation; for consistency, authors adopted the word, accident from HSIS to describe crash.

4 **TWO-STEP CRITERIA SELECTION METHOD**

The following two subsections introduce the two steps, variable pre-selection and random forest, in our proposed Two-Step Criteria Selection Method. The flow of the methodology is illustrated in Figure 1.

**Step 1: Variable Pre-selection**

After integrating the HSIS and the HPMS datasets into one dataset, there are 111 variables in total to be evaluated. 20 variables come from the HSIS, while 91 variables come from HPMS, as shown in Layer 1, Figure 1. A series of data cleaning and checking procedures were considered in the variable pre-selection step, such as the following:

- Eliminate the variables with either empty (i.e., over 90% of N/A) or erroneous data;
- Eliminate the deterministic variables (i.e., with variance close to zero);
- Examine and eliminate the correlated numerical variables; and etc.

After the data cleaning and consistency checking procedures, 111 variables with 35,298 data points are reduced into 27 variables with 8,586 data points. These 27 variables are then categorized into three safety aspects, 12 accident related variables, 10 roadway inventory related variables and 5 traffic related variables (see details listed in Layer 2 in **Figure 1**).
Step 2 Selection: Random Forest

With these 27 pre-selected variables, in Step 2, the Random Forest (RF) machine learning algorithm [25], a popular tree-based regression and classification method, is performed over these 27 variables list Layer 2, Figure 1. The essential idea with using the RF algorithm is to grow an extensive collection of decorrelated trees based on different parts of the same training set and averaging the results. Thus, the algorithm can provide low variance results. Practically, each feature is sampled without replacement according to proportion of its maximum in RF algorithm. Gini index is a common tool to interpretate and rank the feature outcomes from RF. It is defined in Equation (1) and denotes node impurity (the probability of a wrongly classified variable when randomly chosen). Predictors with largest Gini coefficient are chosen to make a binary split on the node.

\[
Gini\ index = \sum_{i}^{n} p_i (1 - p_i) \tag{1}
\]

where \(n_i\) is the number of classes in the target variable and \(p_i\) is the probability of an object being classified to a particular class. In the RF algorithm, the Mean Decrease in Gini index is the weighted average of the predictor's decrease in node impurity. It is effectively a measure of variable importance. A higher Mean Decrease in Gini indicates higher variable importance. In the Step 2 Criteria Selection, the Gini index is computed as in Figure 2.
Figure 2. Step 2 Criteria Selection using Random Forest

The RF algorithm performed with a 94.3% accuracy for the test data during pre-training process. Then, the 27 pre-selected variables are categorized into their aspects (i.e., accident, roadway inventory, and traffic) in Layer 2 and ranked per percentile calculated from Gini Index (i.e., relative importance). Details on the relative importance (i.e., percentile) of each variable is calculated and presented in Figure 3. Lastly, 13 safety criteria are then finalized by choosing those variables with a 50% percentile or above in their aspects. Those safety criteria are implemented for re-training the random forest model. A final model with those key safety criteria reached 95.3% accuracy for the test data.
Figure 3: Variable Importance in Accident, Roadway Inventory and Traffic Related Aspect
RESULTS AND DISCUSSIONS

Two types of comparisons are visualized on the heat maps and discussed in the following section. One compares the performances of identified key safety criteria in each aspect on the existing smart corridor along WA 522. The primary purpose is to evaluate whether those safety criteria describes the characteristics of this existing smart corridor. The other comparison is between potential smart corridors on WA 161, WA 99, WA 202, and the existing one on WA 522. The purpose is to study the similarities and differences between those predicted ones and the existing one on WA 522, and discuss the potentials of them as smart corridors.

Existing Smart Corridor on WA 522

Following the proposed Two-Step Criteria Selection in methodology, the following 13 safety criteria are selected out of 111 variables from HSIS and HPMS database for the State of Washington,

- Accident related aspect: (1) Number of accidents occurred on driveway, (2) Total number of accidents, (3) Number of multi-vehicle accidents, (4) Number of accidents occurred on intersection, (5) Number of accidents in AM Peak, (6) Number of accidents in PM Peak, (7) Number of accidents with property damage only;
- Roadway inventory related aspect: (8) Width of left shoulder, (9) Degree of curve with 28 degree or over, (10) Median width;
- Traffic related aspect: (11) AADT, (12) Number of signalized intersections, and (13) Number of intersections.

The Top 1 and 2 criteria at each aspect are demonstrated in Figure 4, Figure 5 and Figure 6. Those figures evaluate whether these safety criteria identified by RF machine learning are good representatives of a smart corridor.

Figure 4: Heat Maps to Demonstrate the Top 2 Key Criteria in the Accident Related Aspect on the Smart Corridor, WA 522
Figure 4 illustrates the heat maps for the No. 1 and the No. 2 important criterion, from an accident related aspect. Both criteria are representative as they are highlighted (i.e., with an orange/red color) in the existing smart corridor from WA 522 from NE 153rd Street to 83rd Place NE. The number of driveway accidents is relatively more important than the total number of accidents, because driveway accidents identify those segments under the smart corridor the most. While the total number of accidents is large on segments along smart corridor, as well as some segments beside WA 522; the driveway crashes are mostly dense along the existing smart corridor. Therefore, the number of driveway accidents better describes the characteristic of the smart corridor. Overall, both are critical safety criteria to consider, because large crash volumes and/or at specific location (i.e., near driveway) on a signalized high-speed corridor demonstrates the room for the need of V2I applications. Similarly, the heat maps in Figure 5 and Figure 6 illustrate the No. 1 and the No. 2 important criterion from the roadway inventory and traffic related aspects. These criteria are highlighted along WA 522. That is, these criteria well describe the characteristic of the smart corridor. Since poorly designed road inventory or heavy traffic
over the capacity may lead safety issues, the authors believe that these key factors demonstrates
the needs for implementing smart technologies.

**Potential Smart Corridors in WA State Routes**

More than identifying and verifying the key criteria through the characteristics of the existing
corridor on WA 522, four potential smart corridors from three separate state routes (i.e., WA
161, WA 99, WA 202) are predicted. They are selected from a total of 221 state routes in the
Washington state using those 13 identified safety criteria. They are circled in the heat map on
Figure 7. The red color represents a larger probability to be a smart corridor.

![Potential Smart Corridors in WA State Routes](image)

**Figure 7: A Heat Map to Identify Potential Smart Corridors in WA**

These four potential corridors are predicted by the RF algorithm using identified 13
safety criteria. However, they are with a lower selection priority than the existing smart corridor.
Figure 7 maps the locations of those corridors along with the existing one. It is noticeable that
although the potential smart corridor #2, #3 and #4 contain 35, 27 and 23 segments, these
segments are identified separately by the criteria. That is, some segments on the corridor are
around with safety concerns, while some are not. This leads to a lower potential to deploy smart
technology than the existing corridor on WA 522. On the other hand, the potential smart corridor #1 on WA 161 has 85 continuously identified segments. It is almost identical to the existing smart corridor by considering those safety criteria. However, by examining the pre-selected variables, it is noticed that there is a difference brought by the truck percentage. The truck percentage on WA 522 varies from 2% to 8%, whereas it ranges from 2% to 13% on WA 161. For a smart corridor with signalized intersections, truck percentage is an additional factor to consider. It is because a higher truck percentage may minimize the benefit brought by the SPaT message and V2I applications. For example, a connected and autonomous vehicle receives signal timing message and wants to plan its trajectory accordingly to pass the intersection without a stop, but it is limited to speed up or change lanes because of trucks around intersection.

**SUMMARY AND FUTURE STUDY**

This study demonstrated a use of the HSIS dataset to determine key safety criteria for selecting a smart corridor using a machine learning approach, Random Forest. The HSIS contains a rich dataset and it well records data including various variables from many aspects of transportation. In this study, the authors implemented the Random Forest algorithm to finalize 13 safety criteria for selecting a smart corridor out of 111 variables in the HSIS and the HPMS from the State of Washington. Then, by evaluating those criteria with its existing SPaT enabled corridor, the authors believe that those criteria are critical to consider when selecting a smart corridor. These criteria also agree with the guideline from FHWA & NOCoE for selecting SPaT enabled corridor. Lastly, this study predicted potential smart corridors on WA 161, WA 99, WA 202 and discussed their potentials in deploying connected technologies. The safety criteria recommended in this study is generalized and ready to adapt in other states. There are some limitations of this study may lead to future improvements:

- Limited data used (i.e., data in 2015 only)
- Limited area (i.e., only in the State of Washington): The Two-Step Selection Method is adaptable to other states, a more comprehensive study is to use HSIS database in all 8 states.

Nevertheless, as the selection process of a smart corridor is time-consuming and the costs of construction and maintenance are expensive, the 13 safety criteria recommended from this study are important. They are efficient and effective ways for state and local agencies to identify potential smart corridors in their state route network. Lastly, the authors believe that this study is a novel use of the HSIS data and demonstrates a diverse application of the HSIS data with the machine learning technology and the concept of ITS.

**ACKNOWLEDGMENT**

The authors would like to acknowledge Anusha Patel Nujjetty, Manager at the Highway Safety Information System (HSIS) Laboratory, for providing accident data used in this study. In addition, advise from Dr. Lord Dominique as the faculty advisor of this work.
REFERENCES


